Please cite as:

Onghena, P. (2018). Randomization tests or permutation tests? A historical and terminological clarification. In V. Berger (Ed.), *Randomization, masking, and allocation concealment* (pp. 209-227). Boca Raton/FL: Chapman & Hall/CRC Press.

ISBN-13: 978-1-138-03364-1

Randomization tests or permutation tests? A historical and terminological clarification

Patrick Onghena

KU Leuven – University of Leuven, Belgium

Abstract

The terms “randomization test” and “permutation test” are sometimes used interchangeably. However, there are both historical and conceptual reasons for making a clear distinction between the two terms. Using a historical perspective, this chapter emphasizes the contributions made by Edwin Pitman and Bernard Welch to arrive at a coherent theory for randomization and permutation tests. From a conceptual perspective, randomization tests are based on random assignment and permutation tests are based on random sampling. The justification of the randomization test derives from the fact that under the null hypothesis of no treatment effect, the random assignment procedure produces a random shuffle of the responses. The justification of the permutation test derives from the fact that under the null hypothesis of identical distributions, all permutations of the responses are equally likely. It is argued that this terminological distinction is crucial for recognizing the assumptions behind each of the tests and for appreciating the validity of the corresponding inferences.

Randomization tests, as defined in the previous chapter on "Randomization and the randomization test: Two sides of the same coin", are sometimes labeled "permutation tests". Furthermore, they tend to have Fisher's name attached to them, resulting in terms such as "Fisher's randomization test" (Basu, 1980; Jacquez & Jacquez, 2002) or "Fisher's permutation test" (Oden & Wedel, 1975; Soms, 1977). In this chapter, we will provide a clarification of both the history and the current use of these terms. In the historical clarification, we will argue that other statisticians than Fisher played a more decisive role in the development of a coherent theory of the randomization model. In the terminological clarification, we will emphasize that there is a crucial difference between randomization tests and permutation tests. In addition, we will discuss the relationship of randomization tests to other computer-intensive tests and exact tests, and the (ir)relevance of the concept of "exchangeability" for the validity of randomization tests and permutation tests.

HISTORICAL CLARIFICATION

The history of the relatively young scientific discipline of statistics is filled with intriguing scientific puzzles, debates, and controversy (Box, 1978; Hald, 1998, 2007; Reid, 1982; Stigler, 1978, 1986). Current practice can still be informed and inspired by the origins of these puzzles and the genesis of their solutions. In this section, we want to revisit Fisher (1935), Pitman (1937a, 1937b, 1938), and Welch (1937) regarding the development of the randomization model for statistical inference. We will argue that Stigler's law of eponymy also applies for the randomization test: No scientific discovery is named after its original discoverer (Gieryn, 1980).

Fisher (1935)

Fisher (1935) allegedly introduced the randomization test in the third chapter of *The Design of Experiments* in which he re-analyzed Charles Darwin's data on the difference in height between cross- and self-fertilised corn (Alf & Abrahams, 1972; Basu, 1980; Box & Andersen, 1955; Jacquez & Jacquez, 2002; Manly, 2007). However, after closer examination of that chapter, it is obvious that Fisher (1935) did not embrace the randomization model at all. In every section of that chapter, Fisher (1935) firmly endorsed the classical random sampling model:

> The most general statement of our null hypothesis is therefore that the limits to which these two averages tend are equal. *The theory of errors* enables us to test a somewhat more limited hypothesis, which, by wide experience, has been found to be appropriate to the metrical characters of experimental material in biology. The disturbing causes which introduce discrepancies in the means of measurements of similar material are found to produce quantitative effects which conform satisfactorily to a theoretical distribution known as *the normal law of frequency of error*. It is this circumstance that makes it appropriate to choose, as the null hypothesis to be tested, one for which an *exact* statistical criterion is available, namely that the two groups of measurements are *samples drawn from the same normal population*. On the basis of this hypothesis we may proceed to compare the average difference in height, between the crossfertilised and the self-fertilised plants, with such differences as might be expected between these averages, in view of the observed discrepancies between the heights of plants of like origin. (pp. 39-40, italics added)

So from the onset of this chapter, Fisher (1935) was clear about his goal and expressed strong support for “the theory of errors” a.k.a. “the normal law of frequency of error”. This enabled him to use the results of "Student's" *t* test to contest Galton’s analysis of Darwin’s data:

> It has been mentioned that "Student's" *t* test, in conformity with the classical theory of errors, is appropriate to the null hypothesis that the two groups of measurements are *samples drawn from the same normally distributed population*. This is the type of null hypothesis which experimenters, *rightly in the author's opinion*, usually consider it appropriate to test, for reasons not only of practical convenience, but *because the unique properties of the normal distribution make it alone suitable for general application*. (p. 50, italics added)

It is only to ward off critics of "Student's" *t* test regarding the assumption of the normally distributed population that Fisher (1935) used the randomization argument in the last few pages of that chapter. However, he did not loosen the random sampling assumption:

> On the hypothesis that the two series of seeds are random samples from identical populations, and that their sites have been assigned to members of each pair independently at random, the 15 differences of Table 3 would each have occurred with equal frequency with a positive or with a negative sign. (Fisher, 1935, p. 51)

Furthermore, he left no doubt that he considered the randomization argument merely as an auxiliary device to validate the tests based on “the classical theory of errors”:

> The arithmetical procedure of such an examination [i.e., the randomization test] is tedious, and we shall only give the results of its application in order to show the possibility of an independent check on the more expeditious methods in common use ["Student's" *t* test]. (Fisher, 1935, p. 51)

He makes this point more explicit in a 1936 paper:

> Actually, the statistician does not carry out this very simple and very tedious process [i.e., the randomization test], but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method.

For the seventh edition of *The Design of Experiments*, Fisher (1960) appended a short section in the third chapter in which he curbed any enthusiasm regarding “this elementary method”. In this additional section, he simultaneously claimed to have had the scoop and emphatically dismissed the “tests using the physical act of randomization” as only deserving a secondary role:

> In recent years tests using the physical act of randomisation to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of "Non-parametric" tests. Somewhat extravagant claims have often been made on their behalf. The example of this Chapter, published in 1935, was by many years the first of its class. The reader will realise that it was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often,

> wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality. (p. 48)

However, Fisher (1935) cannot be given the scoop for discovering the randomization test because he never abandoned the random sampling idea. In his description of the test on page 51 of *The Design of Experiments*, he makes no conceptual difference between random sampling and random assignment. Instead, the reader is given the impression that any random process validates the use of the Gaussian theory of errors. He thought of "tests using the physical act of randomization" merely as a dispensable tool to confirm the universal validity of classical parametric tests based on random sampling.

At most, the third chapter of *The Design of Experiments* can be considered as an intermediate step in an ongoing discussion of randomization both as a design principle and as a way to validate statistical tests. Considered in Fisher's body of work, it is a provisional argument in a larger project with a different goal (parametric likelihood theory), and Fisher never returned to this argument in his statistical work ever again (Kempthorne, 1976; Rubin, 2005). Furthermore, as noticed by David (2008), Fisher's justification of the Gaussian theory of errors by randomization is predated by a paper of Eden and Yates (1933), prepared at Rothamsted shortly before Fisher left for University College London. Eden and Yates (1933) acknowledge Fisher's "interest and advice", but Fisher (1935) makes no reference to Eden and Yates (1933). As mentioned by Rubin (1990), Imbens and Rubin (2015), Berry, Johnston, and Mielke (2014), and Berry, Mielke, and Johnston (2016), Fisher's ideas about randomization and the repeated-sampling evaluations of randomization distributions were predated by many other publications, most notably by Neyman (1923), "Student" (1923), and Geary (1927).

Pitman (1937a, 1937b, 1938) and Welch (1937)

Edgington and Onghena (2007) pointed out that the first author who proposed the proper randomization model and who presented the randomization test as a test for its own sake was Pitman (1937a, 1937b, 1938). Edwin J. G. Pitman was Professor of Mathematics, working at the remote University of Tasmania in Australia. He "strayed into statistics" after meeting an experimenter at the State Department of Agriculture, R. A. Scott, who brought him some data and statistical analyses from field trials on potatoes, together with a copy of Fisher's *Statistical Methods for Research Workers* (Pitman, 1982; Williams, 1994). Pitman started studying Fisher's publications and verifying his calculations, and immediately understood the broader application of Fisher's validation argument to "significance tests which may be applied to samples from any populations" (Sprent, 1994; Williams, 1994). In a series of papers, he introduced the randomization test for comparing the averages of two independent groups (Pitman, 1937a), the randomization test for the correlation coefficient (Pitman, 1937b), and the randomization test for analysis of variance (Pitman, 1938). He showed how significance tests, confidence intervals, and their approximations can be derived under a random sampling model, but he also provided demonstrations for situations in which "the two samples together form the whole population", which means that there is only random assignment and no population other than the sample itself:

> But the essential point of the method is that we do not have to worry about the populations which we do not know, but only about the sample values which we do know. (Pitman, 1937a, p. 129)

> In particular, the theorem is true when the two samples together form the whole population; it then needs no proof. (Pitman, 1937a, p. 129)

However, Pitman (1937a) made the issue of precedence complicated by humbly giving most of the credit to Fisher (1935):

> The main idea is not new, it seems to be implicit in all Fisher's writings; [Footnote: See, for example, R. A. Fisher, *The Design of Experiments*, p. 50 (Oliver and Boyd).] but perhaps the approach to the subject, frankly starting from the sample and working towards the population instead of the reverse, may be a bit of a novelty. (Pitman, 1937a, p. 119)

Later in his career, Pitman regretted that he wrote that disclaimer:

> I was always dissatisfied with the sentence that I wrote… I wanted to say that I really was doing something new. (Pitman in a letter dated June 1986, quoted in Edgington, 1995, p. 18)

The least one can say, is that it was an exaggeration to refer to "all Fisher's writings".

At about the same time and presumably independent from Pitman, Welch (1937) proposed randomization tests for Randomized Blocks and Latin Square experiments. Bernard L. Welch was working closer to the epicenter of the world's leading statistical research and development: at the University College London, where Egon Pearson was head of the Department of Statistics and Fisher head of the Department of Eugenics (Box, 1978; Mardia, 1990). Later in his career, he would develop the famous Welch *t*-test (as an alternative to Student's *t*-test without assuming equal variances, Welch, 1947), but in his 1937 paper he developed randomization tests along the same lines as Pitman did, and emphasized that the statistical inference resulting from a randomization test was not intended to go beyond the data at hand:

> In experiments in which randomization is performed, the actual arrangement of treatments on the field is one chosen at random from a predetermined set of possible arrangements. In the present paper investigation has been made for Randomized Blocks and Latin Square experiments, into the distribution of the statistic *z*, generated by the application to the observed plot yields of the whole fundamental set of arrangements, assuming as true the "null" hypothesis that the treatments have no differential effect on the plot. (Welch, 1937, p. 47)

> We may make, from the yields of the experiment, a *statistical* inference only about the situation on the particular field of the experiment, e.g. as in the present paper, we may be using our *statistical* method only to test whether all the treatments would have given identical yields on each plot of this particular field. (p. 47, italics in original)

Welch (1937) did not refer to Pitman (1937a, 1937b) but Pitman (1938) referred to Welch (1937). This is remarkable because, as pointed out by Sprent (1994), in the pre-airmail era it took about three months to send letters and papers between London and Hobart. Pitman graciously acknowledged Welch's work in the introduction to his 1938 paper:

> As one of a series, this paper was planned many months ago; but it was not written until June of this year, 1937. It arrived in England just about the time when B. L. Welch's paper on the analysis of variance appeared. While some of its results are anticipated by Welch, the present paper goes deeper into the randomization theory of the simplest type of analysis of variance test. (Pitman, 1938, p. 322)

In the final section, he apologetically added:

> Only the simplest type of analysis of variance test has been discussed in this paper. I had intended another paper to follow, which would deal in the same way with the Latin square arrangement; but this has been dealt with by Welch (5). I may add that Welch's equation (49) on p. 41, giving the variance of *W* for Latin square, agrees with my own result, which was reached by a route quite different from his. In view of the rather heavy algebra involved it seems worth while publishing this confirmation of Welch's result. (Pitman, 1938, p. 335)

Stigler's law of eponymy

So perhaps it is no surprise that Fisher's name became attached to the test: Pitman (1937a) credited Fisher (1935); and Welch (1937) did not refer to Pitman (1937a, 1937b, 1938) because they were working on opposite sides of the globe and in the 1930s only snail mail was available. Furthermore, Fisher was a fervent advocate of randomization, and in his influential publications from a decade earlier, he promoted randomization as one of the basic principles of experimental design, together with blocking and replication (Fisher, 1925, 1926).

Interestingly, another statistical heavyweight, Jerzy Neyman, did something very similar with respect to giving credit to Fisher for the randomization design principle as Edward Pitman did with respect to giving credit to Fisher for developing the test. In his notorious paper, read before the Industrial and Agricultural Research Section of the Royal Statistical Society, Neyman (1935, p. 109) stated:

> Owing to the work of R. A. Fisher, "Student" and their followers, it is hardly possible to add anything essential to the present knowledge concerning local experiments.... One of the most important achievements of the English School is their method of planning field experiments known as the method of Randomized Blocks and Latin Squares.

A few pages later, he added:

> The difficulty has been overcome by the device proposed by R. A. Fisher, which consists in making the η's random variables with mean equal to zero. For this purpose the plots within each block are randomly distributed among the different objects. (Neyman, 1935, p. 112)

Notwithstanding his own 1923 paper on the completely randomized design, Neyman firmly ascribed the introduction of randomization as a physical act, and later as a basis for analysis, to Fisher (Speed, 1990). Neyman's biographer, Constance Reid (1982), recalls:

> On one occasion, when someone perceived him as anticipating the English statistician R. A. Fisher in the use of randomization, he objected strenuously: ". ... I treated theoretically an unrestrictedly randomized agricultural experiment and the randomization was considered

> as a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements." (p. 45)

With all these credits, it might be obvious to also name the test after Fisher. In the end, he is probably one of the most influential and prolific statisticians of the 20th century (Hald, 1998, 2007; Heyde & Seneta, 2001; Lehmann, 2011; Rao, 1992; Salsburg, 2001) so his name may act as a magnet for statistical procedures that are looking for justification, recognition, and respect. In this sense, Fisher's celebrity enjoys the benefit of the Matthew effect as described in the sociology of science (Merton, 1968). Stephen Stigler derived Stigler's law of eponymy from this effect: "No scientific discovery is named after its original discoverer". With Robert Merton (1968) as the true discoverer, this "law" confirms itself (Gieryn, 1980).

For completeness' sake, it should be added that some authors rightly use Pitman's name as the descriptor for the test. For example, references can be found to the "Fisher-Pitman test" (Wilks, 1962), the "Pitman-Welch test" (Feinstein, 1985), and the "Pitman test" (Gibbons, 1986b; Pratt & Gibbons, 1981). These exceptions confirm the rule but usually only refer to a specific design and a specific model. For example, Wilks (1962) used the term "Fisher-Pitman test" to indicate the two-sample test in a random sampling model. Notice also that the so-called "Fisher's Exact Test" for two-by-two contingency tables is in fact a randomization test in a Completely Randomized Design with two treatments and with two categories of responses (Edgington & Onghena, 2007). The origin of Fisher's Exact Test in the *The Design of Experiments* and the resulting controversies are beyond the scope of the present chapter, but the interested reader is referred to Barnard (1945, 1947, 1949), Neyman (1950), Gridgeman (1959), Berkson (1978), Basu (1980), Yates (1984), Agresti (2013), and Bi and Kuesten (2015).

Finally it should be noted that neither Fisher (1935) nor Pitman (1937a, 1937, 1938) or Welch (1937) used the term "randomization test". Fisher referred to "tests using the physical act of randomization". Pitman merely referred to "significance tests which may be applied to samples from any populations", whereas Welch (1937) compared the "Normal Theory" and the "Randomization Theory". According to David (2001), the term "randomization test" makes its first appearance in Box and Andersen's 1955 paper on "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption," in which it is used as a synonym for "permutation test". However, the term randomization test is already prominently present in Moses's 1952 paper on "Non-parametric Statistics for Psychological Research" (the term even appears as a section title on page 133) and in Kempthorne's 1952 classical textbook, "The Design and Analysis of Experiments" (also as a section title, page 128), suggesting an even older (possibly common) source.

TERMINOLOGICAL CLARIFICATION

As is evident from the previous section, the terminology surrounding "tests using the physical act of randomization" has been confusing since their first introduction. Even nowadays still all kinds of terms are used interchangeably to refer to these tests. In this chapter, we want to endorse the proposal first made by Kempthorne and Doerfler (1969) to make an explicit distinction between "randomization tests" and "permutation tests". Randomization tests are based on a random assignment model and permutation tests are based on a random sampling model. The justification of the randomization test derives from the fact that under the null hypothesis of no treatment effect, the random assignment procedure produces just a random shuffle of the responses. The justification of the permutation test derives from the fact that under the null hypothesis of identical distributions, all permutations of the responses are equally likely. This distinction is also followed by Cox and Hinkley (1974), Willmes (1987), Mewhort (2005), Zieffler, Harring, and Long (2011), and Keller (2012). Randomization tests (under the strict random assignment model) are also sometimes called "rerandomization tests" (Brillinger, Jones, & Tukey, 1978; Gabriel, 1979; Gabriel & Hall, 1983; Gabriel & Hsu, 1983; Petrondas & Gabriel, 1983), as an analogy with "resampling tests", but this terminology is rare in the current scientific literature. Furthermore, this terminology might be misleading because it contains a suggestion that a new test procedure is proposed (although it is an "ordinary" randomization test). Finally, the physical act of randomization is a one-time thing, and "rerandomization" is not at stake.

Randomization tests and permutation tests

The strict distinction between "randomization tests" and "permutation tests" based on their different justifications for their validity implies that there is no subset relation between the two types of tests, and that there is even no intersection. However, sometimes authors use the term "permutation tests" to refer to a broader category of tests including randomization tests as a special case (Ernst, 2004; Good, 2005; Lehmann & Romano, 2005; Pesarin & Salmaso, 2010; Scheffé, 1959). This causes no interpretation problems in specific applied research contexts if the models and justifications are also clearly specified. However, generally speaking, having identical terms referring to different procedures and having different terms referring to identical procedures is a recipe for confusion.

Symptomatic for this confusion are the entries in the *Encyclopedia of Statistical Sciences*. In this encyclopedia Gibbons (1986a) defined permutation tests as a particular kind of randomization tests, while Edgington (1986) defined randomization tests as a particular kind of permutation tests. According to Gibbons (1986a) a permutation test is a randomization test whose reference distribution is generated by permutations, and according to Edgington (1986) a randomization test is a permutation test that is based on random assignment. Such contradictory definitions can be avoided by making a clear distinction between randomization tests and permutation tests and linking them firmly to the corresponding random assignment model and random sampling model. One additional advantage of avoiding the term "permutation tests" for the broader category is that we do not have to invent another word for "permutation tests that are no randomization tests" ("permutation tests sensu stricto"). Another additional advantage of avoiding the term "permutation tests" in the definition of randomization tests is that the unfortunate reference to "permutation" as a combinatorial concept disappears. This reference is unfortunate because many random assignment

schedules are not based on permutations but rather on combinations, partitions, or even on random determination of interventions points in randomized single-case phase designs (Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Moir, 1998; Onghena & Edgington, 2005).

Notice that Lehmann and D'Abrera (2006) and Siegel and Castellan (1988) used the term "permutation tests" even in yet another, more restricted way. They used it to make a distinction with nonparametric rank tests. "Permutation tests", according to Lehmann and D'Abrera (2006) and Siegel and Castellan (1988), are applied to the original scores, while nonparametric rank tests are applied to the ranks. Of course, this adds to the confusion. To make things worse, in the first edition of his popular handbook, Siegel (1956) used the term "randomization tests" instead of "permutation tests" for this category of tests. In order to emphasize that the tests are applied to the original scores, other authors used the terms "component-randomization tests" or "observation-randomization tests" as opposed to "rank-randomization tests", which might be more appropriate and less confusing (Alf & Abrahams, 1972; Bradley, 1968; Pratt & Gibbons, 1981; Wilks, 1962).

Randomization tests and computer-intensive tests

The computations involved in performing randomization tests have much in common with other computer-intensive methods, such as the jackknife, the bootstrap, cross-validation, and Monte Carlo tests (Efron, 1979; Efron & Gong, 1983; Efron & Tibshirani, 1991; Manly, 2007; Noreen, 1989; Romano, 1989). This has led some authors to discuss randomization tests in a more general framework of "resampling" or "bootstrapping" (Simon & Bruce, 1991; Westfall & Young, 1993). In such a framework, the only difference between bootstrap significance tests and randomization tests is that with the former method the data are sampled with replacement, while in the latter the data are sampled without replacement (Manly, 2007,; Romano, 1989; Westfall & Young, 1993).

One should, however, not overlook the fact that bootstrap significance tests and randomization tests have different theoretical foundations and different statistical properties. For example, with a randomization test, the Type I error rate is completely under control, while with a bootstrap significance test, the Type I error rate is only under control for large samples in some designs with some test statistics (Manly, 2007; Noreen, 1989; Rasmussen, 1989; ter Braak, 1992). As Efron and Tibshirani (1993) commented:

> The bootstrap distribution was originally called the "combination distribution". It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute. When there *is* something to permute (…) it is a good idea to do so, even if other methods like the bootstrap are also brought to bear. (p. 218)

Randomization tests and exact tests

Randomization tests and permutation tests are sometimes called "exact tests" to the effect that the terms are used as synonyms (Chernick & Friis, 2003; Gacula, Singh, Bi, & Altan, 2009; Higgins, 2017; Krauth, 1988). The "exactness" of the test refers to the exact control of the Type I error rate, and in this sense depends on the kind of assumptions that are involved. "Exact tests" is a term that makes sense in contrast to approximate tests or Monte Carlo tests that only control the Type I error rate

approximately or asymptotically (Mehta & Patel, 1983; Mehta, Patel, & Senchaudhuri, 1988; Mehta, Patel, Senchaudhuri, & Tsiatis, 1994; Senchaudhuri, Mehta, & Patel, 1995).

Although "exact tests" is an appealing term – who wants to use "inexact tests"? – it fails as a synonym for randomization tests or permutation tests. Also parametric tests are exact tests if their assumptions are satisfied (see Weerahandi, 1995, 2004). Furthermore, Dwass (1957) and Hope (1968) developed Monte Carlo versions of randomization tests and permutation tests, which are perfectly valid as tests on their own, but which are, by the above definition, not exact tests. These Monte Carlo versions have become increasingly popular in recent years because of their relative ease of programming as compared to the exhaustive versions in which all possibilities have to be enumerated once and only once. Notice that these Monte Carlo tests are perfectly valid if the observed test statistic is included in the reference set. By contrast, if the "exact *p*-value" is merely estimated by a Monte Carlo procedure without including the observed test statistic in the reference set, as proposed by Senchaudhuri, Mehta, and Patel (1995), then it is no longer a perfectly valid test on its own (Edgington & Onghena, 2007; Onghena & May, 1995; Phipson & Smyth, 2010). If only one hypothesis is to be tested, the consequences of the miscalculation by using the wrong Monte Carlo version are usually negligible, but in a multiple testing situation they may become substantial. As pointed out by Phipson and Smyth (2010):

> In genomic research, however, it is typically the case that many tests are to be conducted. When the number of tests is large, any systematic underestimation of *p*-values can lead to dangerously wrong conclusions at the family-wise level. (p. 4)

Randomization tests and exchangeability

In some publications, "exchangeability" is used as a common foundational assumption of randomization tests and permutation tests (see e.g., Commenges, 2003; Good, 2005; Nichols & Holmes, 2002; Pesarin & Salmaso, 2010; Winkler, Ridgway, Webster, Smith, & Nichols, 2014). Exchangeability is satisfied if the joint distribution of the n observations $y_1, \ldots y_i, \ldots y_n$ is invariant under permutations of the indices. Examples of exchangeable observations are observations from independent, identically distributed (iid) variables and observations from jointly Gaussian distributed variables with identical covariances (Draper, Hodges, Mallows, & Pregibon, 1993; Good, 2002; Greenland & Draper, 2011).

Exchangeability, as a statistical concept, has its origin in Bayesian statistics, in which it is used as a "weaker" assumption than the assumption of iid variables (i.e., all iid variables are exchangeable, but not all exchangeable variables are iid) (Galambos, 1986; Gelman et al. 2014; Koch, 1982; Lindley & Novick, 1981). Although this property of exchangeability is important for the foundation of statistical inference, its immediate relevance for the applied statistician is limited. In his discussion of a paper by Draper et al. (1993), George A. Barnard suggested the term "permutability" as a more apt term for the property of exchangeability. This suggestion reveals the *circulus in probando* if the validity of randomization tests and permutation tests is grounded in exchangeability: the observations can be permuted because they are exchangeable. This means that the observations can be permuted because the observations can be permuted. Furthermore, the validity of randomization tests is not based on the exchangeability of the observations as such, but rather on the exchangeability of all elements in the reference set. As mentioned before, many random assignment schedules do not

imply permutations of the observations (Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Moir, 1998; Onghena & Edgington, 2005).

For the validity of randomization tests and permutation tests a reference to actual random assignment or actual random sampling suffices. This reference connects our statistical arithmetic to real world phenomena and actions. Random assignment or random sampling schedules show in which way the experimental units −and therefore also the observations− are exchangeable, but the concept of "exchangeability" is not needed to provide this foundation. It is just rewording the obvious. For data collected without random assignment and without random sampling, any "observational inference" uses an *as if* modus; this means that inferences are made *as if* there was actual random assignment or actual random sampling (Dekkers, 2011; Kempthorne, 1979; Vandenbroucke, 2004). Although the relevance of the concept of exchangeability for the validation of randomization tests and permutation tests is limited, it should be acknowledged that the concept is theoretically important, for example to demonstrate the common foundation of Bayesian and permutational inference and the development of new statistical techniques in the absence of random assignment or random sampling (Draper et al., 1993; Good, 2002; Hutson & Wilding, 2012).

CONCLUSION

In this chapter, we clarified some terms associated with significance tests based on random assignment and we shed some light on the historical development of these tests. Clarification of terms is important for current practice and dissemination. Clarification of the historical development is important to give proper credit to the researchers involved in laying the foundations of the modern use of these statistical techniques and to understand and appreciate the basic questions and intricacies that they were focusing on.

In sum, Fisher (1935) was very successful in his advocacy for the use of randomization in experimental research and for its importance for the validity of statistical tests, but he never systematically elaborated or even showed any interest for the theory behind the significance tests based on random assignment alone. Pitman (1937a, 1937b, 1938) and Welch (1937) picked up that gauntlet and proposed randomization tests as statistical techniques for their own sake.

For the modern use of these statistical techniques it is crucial to acknowledge the difference between a random assignment model and a random sampling model. This is accomplished most conveniently by following the terminology of Kempthorne and Doerfler (1969): significance tests based on the random assignment model are called "randomization tests". Other computer-intensive tests, such as permutation tests and bootstrap tests, are based on the random sampling model and eventually other accompanying assumptions. This difference is important to know exactly what kind of assumptions are involved and whether conditional or unconditional power calculations are most appropriate in the planning stage of the clinical trial (Keller, 2012; Kempthorne & Doerfler, 1969; Pesarin & De Martini, 2002). Finally, the difference between the two kinds of models is crucial to understand the kind of inference that is aimed at (inference to a causal proposition or inference to the sampled population/generalization), and consequently to understand the kind of criteria that have to be used to judge their validity (internal versus external validity) (Shadish, Cook, & Campbell, 2002).

For Tukey (1993) statistical inference in clinical trials based on the actual random assignment constitutes the “platinum standard” of data analysis:

> "platinum standard" describes the ultimate of tightness: probability statements that depend on only exactly how the trial was conducted--not at all on assumptions, such as shapes of probability distributions, that are less (often far less) than completely verifiable. (pp. 266–267)

To recognize this platinum standard, we recommend using the right term unambiguously: “randomization test” or, if you want to pay tribute to the founding fathers: “Pitman-Welsh randomization test”.

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Alf, E. F., & Abrahams, N. M. (1972). Comment on component-randomization tests. *Psychological Bulletin, 77,* 223–224.

Barnard, G. A. (1945). A new test for 2×2 tables. *Nature, 156*(3954), 177.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika, 34,* 123–138.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Series B, 11,* 115–149.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association, 75,* 575–582.

Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistic Planning and Inference, 2,* 27–42.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2014). *A chronicle of permutation statistical methods: 1920–2000, and beyond.* New York, NY: Springer.

Berry, K. J., Mielke, P. W., & Johnston, J. E. (2016). *Permutation statistical methods: An integrated approach.* New York, NY: Springer.

Bi, J., & Kuesten, C. (2015). Revisiting Fisher's 'Lady Tasting Tea' from a perspective of sensory discrimination testing. *Food Quality and Preference, 43,* 47–52.

Box, G. E. P., & Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption, *Journal of the Royal Statistical Society, Series B, 17,* 1–34.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist.* New York, NY: Wiley.

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice-Hall.

Brillinger, D. R., Jones, L. V., & Tukey J. W. (1978). *The Management of Weather Resources II: The Role of Statistics in Weather Resources Management.* Washington, DC: U.S. Government Printing Office.

Chernick, M. R., & Friis, R. H. (2003). *Introductory biostatistics for the health sciences: Modern applications including bootstrap.* New York, NY: Wiley.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Nonparametric Statistics, 15,* 171–185.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London, UK: Chapman and Hall.

David, H. A. (2001). First (?) occurrence of common terms in statistics and probability. In H. A. David & A. W. F. Edwards (Eds.), *Annotated Readings in the History of Statistics* (pp. 219–228 and Appendix B). New York, NY: Springer.

David, H. A. (2008). The beginnings of randomization tests. *The American Statistician, 62,* 70–72.

Dekkers, O. M. (2011). On causation in therapeutic research: Observational studies, randomized experiments and instrumental variable analysis. *Preventive Medicine, 53,* 239–241.

Draper, D., Hodges, J. S., Mallows, C. L., & Pregibon, D. (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A, 156,* 9–28.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics, 28,* 181–187.

Eden, T., & Yates, F. (1933). On the validity of Fisher's *z* test when applied to an actual example of non-normal data. *The Journal of Agricultural Science, 23,* 6–17.

Edgington, E. S. (1986). Randomization tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 7* (pp. 530–538). New York, NY: Wiley.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York, NY: Marcel Dekker.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7,* 1–26.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37,* 36–48.

Efron, B., & Tibshirani, R. J. (1991). Statistical data analysis in the computer age. *Science, 253*(5018), 390–395.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Boca Raton, FL: Chapman & Hall/CRC.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19,* 676–685.

Feinstein, A. R. (1985). *Clinical epidemiology: The architecture of clinical research.* Philadelphia, PA: W. B. Saunders Company.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503–513.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1936). "The Coefficient of Racial Likeness" and the future of craniometry. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 66,* 57–63.

Fisher, R. A. (1960). *The design of experiments* (7th ed.). Edinburgh, UK: Oliver & Boyd.

Gabriel, K. R. (1979). Some statistical issues in weather experimentation. *Communications in Statistics: Theory and Methods, 8,* 975–1015.

Gabriel, K. R., & Hall, W. J. (1983). Rerandomization inference on regression and shift effects: Computationally feasible methods. *Journal of the American Statistical Association, 78,* 827–836.

Gabriel, K. R., & Hsu, C. F. (1983). Power studies of rerandomization tests, with application to weather modification experiments. *Journal of the American Statistical Association, 78,* 766–775.

Gacula, M. C., Jr., Singh, J., Bi, J., & Altan, S. (2009). *Statistical methods in food and consumer research.* Amsterdam, The Netherlands: Elsevier.

Galambos, J. (1986). Exchangeability. In S. Kotz and N. L. Johnson (Eds). *Encyclopedia of statistical sciences, Vol. 3* (2nd ed.) (pp. 2136–2140). New York, NY: Wiley.

Geary, R. C. (1927). Some properties of correlation and regression in a limited universe. *Metron: International Journal of Statistics, 7,* 83–119.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gibbons, J. D. (1986a). Permutation tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 6* (pp. 690). New York, NY: Wiley.

Gibbons, J. D. (1986b). Pitman tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 6* (pp. 740–743). New York, NY: Wiley.

Gieryn, T. F. (Ed.) (1980). *Science and social structure: A festschrift for Robert K. Merton.* New York, NY: Academy of Sciences.

Good, P. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods, 1,* 243–247.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York, NY: Springer.

Greenland, S., & Draper, D. (2011). Exchangeability. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 474–476). Heidelberg, Germany: Springer.

Gridgeman, N. T. (1959). The lady tasting tea, and allied topics. *Journal of the American Statistical Association, 54,* 776–783.

Hald, A. (1998). *A history of mathematical statistics.* New York, NY: Wiley.

Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935.* New York, NY: Springer.

Heyde, C. C., & Seneta, E. (2001). *Statisticians of the centuries.* New York, NY: Springer.

Higgins, J. J. (2017). *Introduction to modern nonparametric statistics* [Online]. Content Technologies Inc.

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B, 30,* 582–598.

Hutson, A. D., & Wilding, G. E. (2012). Maintaining the exchangeability assumption for a two-group permutation test in the non-randomized setting. *Journal of Applied Statistics, 39,* 1593–1603.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences: An introduction.* New York, NY: Cambridge University Press.

Jacquez, J. A., & Jacquez, G. M. (2002). Fisher's randomization test and Darwin's data: A footnote to the history of statistics. *Mathematical Biosciences, 180,* 23–28.

Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika, 77,* 324–338.

Kempthorne, O. (1952). *The design and analysis of experiments.* New York, NY: Wiley.

Kempthorne, O. (1976). Comment on "On Rereading R. A. Fisher" by L. J. Savage. *Annals of Statistics, 4,* 495–497.

Kempthorne, O. (1979). Sampling inference, experimental inference and observation inference. *Sankhyā: The Indian Journal of Statistics, Series B, 40,* 115–145

Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika, 56,* 231–248.

Koch, G. (Ed.) (1982). *Exchangeability in probability and statistics.* Amsterdam, The Netherlands: North Holland.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124–144.

Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach.* Amsterdam, The Netherlands: Elsevier.

Lehmann, E. L., & D'Abrera, H. J. M. (2006). *Nonparametrics: Statistical methods based on ranks* (rev. ed.). New York, NY: Springer.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics.* New York, NY: Springer.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of Statistics, 9,* 45–58.

Manly, B. F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Mardia, K. V. (1990). Obituary: Professor B. L. Welch. *Journal of the Royal Statistical Society, Series A, 153,* 253–254.

Mehta, C. R., & Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in r x c contingency tables. *Journal of the American Statistical Association, 78,* 427–434.

Mehta, C. R., Patel, N. R., & Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association,* 83, 999–1005.

Mehta, C. R., Patel, N. R., Senchaudhuri, P., & Tsiatis, A. A. (1994). Exact permutational tests for group-sequential clinical trials. *Biometrics, 50,* 1042–1053.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159(3810),* 56–63.

Mewhort, D. J. K. (2005). A comparison of the randomization test with the *F* test when error is skewed. *Behavior Research Methods, 37,* 426–435.

Moir, R. (1998). A Monte Carlo analysis of the Fisher randomization technique: Reviving randomization for experimental economists. *Experimental Economics, 1,* 87–100.

Moses, L. E. (1952). Non-parametric statistics for psychological research. *Psychological Bulletin, 49,* 122–143.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Roczniki Nauk Rolniczych Tom X*, 1–51 (Annals of Agricultural Sciences) Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, Statistical Science, 1990, Vol. 5, No. 4,465–480.

Neyman, J., with cooperation of K. Iwaskiewicz and St. Kolodziejczyk (1935). Statistical problems in agricultural experimentation (with discussion). *Journal of the Royal Statistical Society, Series B, 2,* 107–180.

Neyman, J. (1950). *First course in probability and statistics.* New York, NY: Henry Holt.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15,* 1–25.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses: An introduction.* New York, NY: Wiley.

Oden, A., & Wedel, H. (1975). Arguments for Fisher's permutation test. *The Annals of Statistics, 3,* 518–520.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: single-case design and analysis. *The Clinical Journal of Pain, 21,* 56–68.

Onghena, P., & May, R. B. (1995). Pitfalls in computing and interpreting randomization test *p* values: A commentary on Chen and Dunlap. *Behavior Research Methods, Instruments, & Computers, 27,* 408–411.

Pesarin, F., & De Martini, D. (2002). On unbiasedness and power of permutation tests. *Metron: International Journal of Statistics, 60,* 3–19.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software.* Chichester, UK: Wiley.

Petrondas, D. A., & Gabriel, K. R. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association, 78,* 949–957.

Phipson, B., & Smyth, G. K. (2010). Permutation *p*-values should never be zero: Calculating exact *p*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology, 9*(1) [Online] https://doi.org/10.2202/1544-6115.1585.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society, Series B, 4,* 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations II: The correlation coefficient. *Journal of the Royal Statistical Society, Series B, 4,* 225–232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations III: The analysis of variance test. *Biometrika, 29,* 322–335.

Pitman, E. J. G. (1982). Reminiscences of a mathematician who strayed into statistics. In J. Gani (Ed.), *The making of statisticians* (pp. 112–125). New York, NY: Springer.

Pratt, J. W., & Gibbons, J. D. (1981). *Concepts of nonparametric theory.* New York, NY: Springer.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science, 7,* 34–48.

Rasmussen, J. L. (1989). Computer-intensive correlational analysis: Bootstrap and approximate randomization techniques. *British Journal of Mathematical and Statistical Psychology, 42,* 103–111.

Reid, C. (1982). *Neyman from life.* New York, NY: Springer.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics, 17,* 141–159.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science, 5,* 472–480.

Rubin, D. B. (2005). Causal Inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York, NY: W. H. Freeman.

Scheffé, H. (1959). *The analysis of variance.* New York, NY: Wiley.

Senchaudhuri, P., Mehta, C. R., & Patel, N. R. (1995). Estimating exact *p* values by the method of control variates or Monte Carlo rescue. *Journal of the American Statistical Association, 90,* 640–648.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw-Hill.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York, NY: McGraw-Hill.

Simon, J. L., & Bruce, P. (1991). Resampling: A tool for everyday statistical work. *Chance, 4,* 22–32.

Soms, A. P. (1977). An algorithm for the discrete Fisher's permutation test. *Journal of the American Statistical Association, 72,* 662–664.

Speed, T. P. (1990). Introductory remarks on Neyman (1923). *Statistical Science, 5,* 463–464.

Sprent, P. (1994). E. J. G. Pitman, 1897-1993. *Journal of the Royal Statistical Society, Series A, 157,* 153–154.

Stigler, S. M. (1978). Mathematical statistics in the early states. *Annals of Statistics, 6,* 239–265.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

"Student" (1923). On testing varieties of cereals. *Biometrika, 15,* 271–293.

ter Braak, C. J. F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In K.-H. Jöckel, Rothe, G., & Sendler, W. (Eds.), *Bootstrapping and related resampling techniques* (pp. 79-86). Berlin, Germany: Springer.

Tukey, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials, 14,* 266–285.

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomised trials? *Lancet, 363,* 1728–1731.

Weerahandi, S. (1995). *Exact statistical methods for data analysis.* New York, NY: Springer.

Weerahandi, S. (2004). *Generalized inference in repeated measures: Exact methods in MANOVA and mixed models.* New York, NY: Wiley.

Welch, B. L. (1937). On the *z*-test in randomized blocks and Latin squares. *Biometrika, 29,* 21–52.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34,* 28–35.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for* p-*value adjustment.* New York, NY: Wiley.

Wilks, S. S. (1962). *Mathematical statistics.* New York, NY: Wiley.

Williams, E. J. (1994). Edwin James George Pitman 1897-1993. *Historical Records of Australian Science, 10,* 2.

Willmes, K. (1987). *Beiträge zu Theorie und Anwendung von Permutationstests in der uni- und multivariaten Datenanalyse.* Unpublished doctoral dissertation, Mathematische-Naturwissenschaftliche Fakultät, Universität Trier.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage, 92,* 381–397.

Yates, F. (1984). Tests of significance for 2 x 2 contingency tables (with discussion). *Journal of the Royal Statistical Society, Series A, 147,* 426–463.

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R.* Hoboken, NJ: Wiley.